Int J Life Cycle Assess (2012) 17:499–503
DOI 10.1007/s11367-011-0371-x

Commentary on issues in data quality analysis in life cycle assessment

Joyce Smith Cooper · Ezra Kahn

Received: 8 September 2011 / Accepted: 12 December 2011 / Published online: 18 January 2012

Abstract
Purpose For compliance with the ISO standard 14044, comparative life cycle assessments are required to address data quality for time-related coverage, geographic coverage, technology coverage, precision, completeness, representativeness, consistency, reproducibility, sources of the data and uncertainty of the information. As the community of practitioners and data developers grows, the purpose of this commentary is to initiate discussion of current issues and opportunities for improvement in data quality analysis.
Methods Commonly applied data quality analysis methods are described as ranging from the collection of only qualitative information to the assignment of numeric scores. Common interpretations of data quality information are described as ranging from comparison in raw form to contribution and sensitivity analysis results, combination into an aggregate/multiaspect score, or use to infer data uncertainty. Method strengths and issues are described.
Results The strengths of current data quality analysis methods lie in the consideration of the data quality aspects specified by the ISO standards and in the differentiation of low and high data quality. Weaknesses, however, lie in unrepeatable scoring criteria, aggregation of data quality information in a way that is difficult to interpret or misinterpreted and the use of data quality information in the estimation of uncertainty with no basis for accuracy.
Conclusion It is found that among commonly applied methods there exists a need for improved repeatability and interpretability. When combined with emerging efforts to provide reliable uncertainty data to support the use of data quality information with contribution and sensitivity analysis results and efforts that have improved consideration of completeness, the future of data quality analysis promises substantial contribution to the field.

Keywords Data quality · Data quality analysis · Life cycle assessment · Uncertainty

1 Introduction

For life cycle assessment (LCA), the ISO14044 (International Standards Organization ISO, 2006) defines *data quality* as "characteristics of data that relate to their ability to satisfy stated requirements." The standard states that "where a study is intended to be used in comparative assertions intended to be disclosed to the public, the [following] data quality requirements" *shall* be addressed: time-related coverage, geographic coverage, technology coverage, precision, completeness, representativeness, consistency, reproducibility, sources of the data and uncertainty of the information. Among the aspects listed, consistency is applicable to life cycle inventories as opposed to unit process data[1]. Also, completeness is applicable at the level of the unit process and is intended to "ensure that all relevant information and data needed for interpretation is available and complete." The standard subsequently provides information for the treatment of missing data as further guidance for assessing completeness. The remaining eight

Responsible editor: Ralph K. Rosenbaum

J. S. Cooper (✉) · E. Kahn
Design for Environment Laboratory, University of Washington,
Box 352700, Seattle, WA 98195-2600, USA
e-mail: cooperjs@u.washington.edu

[1] The standard defines consistency check as the "process of verifying that assumptions, methods and data are consistently applied throughout the study and are in accordance with the goal and scope definition performed before conclusions are reached" with the objective of determining whether the assumptions, methods and data are consistent with the goal and scope.

requirements are at the level of each flow within a unit process data set.

Whereas data quality analysis methods tend to address most or all of the ISO aspects, execution and the use of results in the interpretation phase of LCA varies. Methods of execution range from the collection of only qualitative information to the assignment of numeric scores. For interpretation, data quality analysis results have been compared in raw form to contribution and sensitivity analysis results, combined into an aggregate/multiaspect score, or used to infer data uncertainty. As the community of practitioners and data developers grows, a discussion of issues and opportunities for improvement in data quality analysis is timely.

2 Data quality analysis methods

In the U.S. LCI Database Project Development Guidelines (Athena™ Sustainable Materials Institute, 2004), data quality is to be described on the bases of data age, source and collection method; data representativeness (e.g., the percentage of total production represented by a sample), averaging methods, and approaches for protecting competition-sensitive company-specific information; methods used to estimate or justify missing data; and information for replicating key assumptions or methodological choices. Uncertainty information for primary (measured) data is to include all available information as either descriptive statistics and associated distribution shapes or an estimate of the uncertainty of the data based on sources or expert judgment. Uncertainty information for secondary (estimated) data is to include all information available from the source that characterized the uncertainty of the data. Thus, in the U.S. LCI Database Project, no differentiation is made between data of lower or higher quality, for example, data with descriptive statistics are treated in the same way as a single data point from a personal communication. An updated set of guidelines is in preparation for the project, the draft versions including options for such differentiation.

One method of differentiating between data of lower and higher qualities that is popular is data quality scoring systems, sometimes called pedigree matrices, and notably including that used by the European Reference Life Cycle Data System[2] (ELCD) through the International Reference Life Cycle Data System (ILCD) data format and used in the ecoinvent database[3] through the ecospold formats. In ILCD (European Commission–Joint Research Centre–Institute for Environment and Sustainability, 2010), data quality scores rank flow data on the bases of six indicators (technological representativeness, geographical representativeness, time-related representativeness, completeness, precision/uncertainty, and methodological appropriateness and consistency) with scores from 0 to 5 (with a score of 1 representing the highest data quality, 5 the lowest and a score of 0 representing data quality that is deemed not applicable) assigned to each. ILCD performs a "missing data" completeness check outside of the data quality scoring, provides guidance on how to judge the completeness of both inventory and impact cutoff and requires that final completeness be reported as the "% degree of completeness/cutoff." ILCD also uses consistency and reproducibility checks outside of the data quality scoring method and targeted at the combination of unit process data into inventories and impact assessments.

The ILCD data quality scoring system is, in some ways, like that supported by ecoinvent v1 and v2 (Frischknecht et al., 2007). For example, ecoinvent v1 and v2 also rank indicator flows, albeit on somewhat different bases when compared to ILCD (using reliability, completeness, temporal correlation, geographical correlation, further technical correlation and sample size) and assigns scores from 1 to 5 (with a score of 1 representing the highest data quality, 5 the lowest and thus, without the not applicable option). Also, like the ILCD method, ecoinvent v1 and v2 perform a "missing data" completeness check outside of the data quality scoring.

Although having the benefit of differentiating between data of lower and higher qualities, inspection of the ILCD and ecoinvent v1 and v2 scoring methods reveals subjectivity that ultimately hampers repeatability. For the ILCD method, subjectivity exists in how scores are assigned, most notably related to interpretation of the phrases "high degree" and "sufficient degree." For the ecoinvent v1 and v2 method, repeatability is, in part, explored by Weidema (1998) with notable subjectivity in the differentiation of "personal information by letter, fax or e-mail based on an estimate by an industrial expert" and "an estimate by an industrial expert" (in the reliability category) and in the interpretation of the phrases "an *adequate* period to even out normal fluctuations" and "*shorter* periods" (in the completeness category) and the terms "related" and "different" (in the technology correlation category). Also, in the sample size category, it appeared that only continuous processes were considered and that sample sizes of 100, 20, 10, and three were assumed to differently represent a population irrespective of the rate of production.

Thus, in the cases of the ecoinvent v1 and v2 and the ILCD scoring methods, much of the benefit scoring provides is lost in the potential for inconsistent application of the scoring method. In ecoinvent v3 (Weidema et al., 2011), data quality scores are developed on a modified five-point scale when compared to the earlier versions. Modifications include (a) moving consideration of sample size from the scoring matrix to the specification of the percent of data sampled out of the total of the

[2] Available at http://lca.jrc.ec.europa.eu/lcainfohub/datasetArea.vm.
[3] Available at http://www.ecoinvent.ch/.

activity as a part of modeling and validation and (b) rewording the reliability indicator description to remove the similarity between points in the scale. However, reproducibility issues remain in the completeness and further technology correlation categories.

To address the issue of data quality analysis reproducibility, a method has been proposed pursuant to the development of data for the US Department of Agriculture's LCA Digital Commons, an open access database and toolset (see http://www.lcacommons.gov/). At the flow level, the Digital Commons intends to use a two-tiered scale presented in Table 1. Using these criteria, each flow is assigned a score of *A* or *B* for each of the seven flow data quality indicators. A score of *A* indicates that the flow data are of higher quality (meeting the requirements in Table 1), and a score of *B* indicates that the unit process data are of lower quality (not meeting the requirements in Table 1). Preliminary use of the Digital Commons method appears to improve reproducibility, albeit at the expense of a finer data quality scale that could be useful during interpretation.

3 Using data quality analysis results in LCA interpretation

For interpretation, the issue at hand is propagation of the data quality information or scores through to the inventory and impact assessment results. Weidema and Wesnæs (1996) note that if it is assumed that, e.g., an inventory flow receives the quality of the lowest quality unit process data point on which it is based (i.e., the quality of the combined data can never be better than the quality of the data contributing to the total), the LCA results are very often entirely of low quality. They note that the result is not very informative, revealing only that a data quality problem has occurred somewhere in the study but not where nor how serious the problem is.

In response, some mathematically aggregate data quality results to produce an overall multiaspect score, first, at the unit process level and then through the inventory and impact assessments. For example, ILCD data quality scores are used to develop an overall measure of data quality as an average score excluding those receiving a score of zero. This method implicitly values all data quality aspects

Table 1 LCA Digital Commons flow data quality scoring criteria

Category	Requirements for a data quality score of *A*
1. Reliability and reproducibility	The flow data were based on measurements using a specified and standardized measurement method OR
	The flow data were estimated using methods and data described in specified archival or other consistently publically available sources.
2. Flow data completeness	The flow data were collected over at least 3 years for agricultural (crop, livestock, forest and range) processes or other processes in which the data point varies for uncontrolled annual conditions (e.g., weather) AND
	The flow data balance the mass and energy in and out of the unit process.[a]
3. Temporal coverage	The flow data represent operations that occurred between the unit process start and end dates without forecasting.
4. Geographical coverage	The flow data represent operations that occurred within the location of the unit process including nonagricultural process data that have been adapted to reflect logistics and market shares[b] for the unit process location.
5.Technological coverage	The flow data represent the process(es) and/or material(s) specified without surrogacy or aggregation with other technologies.
6. Uncertainty	The flow data either include estimates of the first quartile, mean, median and third quartile values OR data or probability distribution from which these values can be estimated.
7. Precision[c]	The relative standard error of the flow data is less than or equal to 25% OR
	The interquartile range divided by the median is less than or equal to 50% OR
	For a triangular distribution, the minimum flow data value is ≥75%, and maximum flow data value is ≤125% of the most likely value OR
	For a uniform distribution, the minimum flow data value is ≥75%, and maximum flow data value is ≤125% of the average of the minimum and maximum values.

[a] An incomplete mass balance may represent either an incomplete unit process or an incomplete set of emissions factors, or both. In the case of a score of *B*, e.g., for an incomplete set of emissions factors, the data quality analysis serves to highlight an opportunity to improve data quality through methodological or documentation improvement

[b] Market shares, sometimes called mixer processes in LCA, reflect the technologies used in local markets. For example, market shares are used to represent the mix of technologies used in regional electricity generation (the percentage of coal, natural gas, nuclear, etc. per kilowatt hour) and the mix of waste management technologies (landfilling, waste-to-energy, etc.) locally available

[c] In the precision category, percentages are intended to represent quartiles, as frequently used in descriptive statistics to represent a fourth of the population being sampled. Note also that for unit processes that balance in category 2, precision will apply as propagated to flows on both sides of the balance

equally, and it is not clear that combining qualitative scores in this way will not lead to misinterpretation (e.g., is an ILCD score of 2 actually twice as good as a score of 4?). Although such an "overall" data quality score might be desirable (e.g., for assessments intended to compare entire life cycles), there is no basis for concluding relative superiority among results across a large number of processes and flows.

Instead of preparing an aggregate score, some consider data quality information as a contributor to the uncertainty of the data, e.g., using the data quality scores to estimate the 'additional' uncertainty resulting from lower data quality (Weidema and Wesnæs, 1996). This path is followed by ecoinvent using its data quality scores to estimate the "square of the geometric standard deviation (95% interval—SDg95)". For the SDg95, a log-normal distribution is assumed, and uncertainty factors associated with each quality indicator based on "expert judgment" are used to calculate the SDg95 to be used in uncertainty analysis. Such transformations of data quality scores to probability distributions are explored by Lloyd and Ries (2007), who warn that unless distribution forms and parameters are defined for specific scores and parameter contributions, there is no basis for their accuracy. Within this context, the ecoinvent Center has commissioned an empirical study to validate and revise the basic uncertainty factors used in the estimation of the SDg95, which will not be issued from quality considerations but instead from expert estimates of variability and imprecision for specific categories of inventory flows. The effort is in recognition of a tendency for the current data quality score-based method to underestimate "real" uncertainty (Weidema et al., 2011).

All databases including the Digital Commons will benefit from these ongoing efforts to provide reliable uncertainty data to be used in estimating inventory uncertainty and performing sensitivity analysis. Given this, we argue that the most transparent and universally applicable use of data quality analysis results is to simply compare them to contribution and sensitivity analysis results: when data of low quality influence the LCA results, higher quality data should be sought or substantial discussion should be provided. Thus, the Digital Commons intends to allow the data quality propagation issue to occur at the unit process, life cycle stage, inventory and impact assessment levels. Given only two tiers of data quality scores, it is very likely that for any given LCA, the data quality will be consistently low at the higher levels but that inspection of the underlying data will reveal critical instances of the lower data quality. What becomes important for a transparent interpretation is to compare aspect-specific data quality information with contribution and sensitivity analyses performed at the scaled flow level (i.e., the quantity in the final inventory) to reveal the critical flows of low quality. Clearly for inventories with thousands of flows, this is only computationally possible with careful and automated data management. However, such computations are compatible with and extractable from those required for uncertainty analysis (e.g., using a Monte Carlo simulation).

Consider for example an LCA of livestock production in which the amount of lime used in the production of feed corn does not meet the Digital Commons data quality criteria for flow data completeness (e.g., data were collected for a single year). The data quality score of *B* applies to: (1) flows in corn production that depend on the amount of lime used (e.g., the transport of lime to the field, the use of lime application equipment, lime-related CO_2 emissions to air at the farm, etc.); (2) all flows in the life cycles of lime production, transport, and application which are scaled to the amount of lime applied; (3) all aggregate flows in the livestock production life cycle that are identified in (1) and (2); and (4) all impact indicators estimated using flows identified in item (3). Whereas the scores of *B* for flow data completeness are largely uninformative at the livestock production and impact assessment levels (items (3) and (4)), they are highly informative as related to lime-related emissions in corn production (within item (1)) and all flows for the life cycles of lime production, transport, and application (item (2)). Thus, a transparent interpretation would use contribution and sensitivity analyses to track the role of every flow in an inventory dependent on the amount of lime applied, quantified as the scaled flow values in each analysis and named by relevant unit processes and data quality scores (e.g., the contribution to the livestock life cycle of CO_2 emissions from lime kilns as based on incomplete data).

4 Discussion

The strengths of current data quality analysis methods lie in the consideration of the data quality aspects specified by the ISO standards and in the differentiation of low and high data quality. Weaknesses, however, lie in unrepeatable scoring criteria, aggregation of data quality information in a way that is difficult to interpret or misinterpreted, and the use of data quality information in the estimation of uncertainty with no basis for accuracy.

The intent of the Digital Commons data quality method is to allow easier and more consistent scoring among practitioners while still communicating what is good and bad about the data. Further, it seems it is less likely that an A/B system will be numerically misinterpreted and/or misused (e.g., average total scores or other numeric combinations without meaning would not be estimated). However, the system is being built on the premise that data quality and uncertainty should be analyzed separately, noting that data quality is not a part of data uncertainty but data uncertainty is a part of data quality. For the system to work, uncertainty

data must become more available than it is today, and contribution and sensitivity analyses must be performed at the scaled flow level.

In the short term, in the absence of comprehensive uncertainty data (e.g., data with estimates of sampling error), it seems the use of a repeatable scoring system for comparison to contribution analysis results is all that is reliably possible. However, this does not facilitate comparative assessments that require an investigation of uncertainty to ensure the systems compared are actually different. In the longer term, data in LCA databases and software can and should be populated with information to facilitate a range of repeatable data quality analysis methods as well as uncertainty data developed using standard and well-established statistical methods. Presuming that perfect uncertainty data will never be available (i.e., that we will continue to use engineering and economic models to represent emerging technologies and markets), methods for the estimation of uncertainty as reviewed by Lloyd and Ries (2007) should be further investigated to fill data gaps.

We recognize the need to plan for compatibility with existing and future LCA software and databases (e.g., ecoinvent, ILCD and the US Database) and hope to initiate a dialogue to develop universally applicable/compatible methods. In the medium term, the Digital Commons team will investigate crosswalking data quality information among databases, and note that the addition of separate data fields in ecoinvent v3 for data quality information will assist in this process. As such, we invite and encourage the community at large to comment on and contribute to this effort.

Beyond all of this, it is very important to note that ecoinvent v3 offers substantial improvements in the definition of completeness (Weidema et al., 2011). For example, monetary, mass and water balances are enforced for all activities and flows (with the exception of nuclear reactions). Related completeness efforts in the LCA community are focusing on the development of "product category rules," as described by ISO 14025, in which relevant flows are defined for a specific category of products (thus, describing what ultimately would be balanced). Further, work on data quality methods is active, for example, as part of an effort under the umbrella of the UNEP/SETAC Life Cycle Initiative, as a continuation of that recently published (see United Nations Environment Programme, 2011). When combined with efforts to improve data quality analysis repeatability, the future of data quality analysis promises substantial contribution to the field.

Acknowledgement This research was funded by the United States Department of Agriculture (USDA) National Agricultural Library (agreement number 58-8201-0-149) and is part of the development of the LCA Digital Commons.

References

Athena™ Sustainable Materials Institute (2004) US LCI Database Project Development Guidelines. US Department of Energy, National Renewable Energy Laboratory. Retrieved from www.nrel.gov/lci/docs/dataguidelinesfinalrpt1-13-04.doc

European Commission–Joint Research Centre–Institute for Environment and Sustainability (2010) International Reference Life Cycle Data System (ILCD) handbook—general guide for life cycle assessment—detailed guidance. Publications Office of the European Union, Luxembourg

Frischknecht R, Jungbluth N, Althaus H-J, Doka G, Heck T, Hellweg S, Hischier R et al (2007) Overview and methodology. Swiss Centre for Life Cycle Inventories, Dübendorf, Switzerland

International Standards Organization (ISO) (2006) Life cycle assessment—requirements and guidelines

United Nations Environment Programme (2011) Global guidance principles for life cycle assessment databases. A basis for greener processes and products, available at http://www.unep.org/pdf/Global-Guidance-Principles-for-LCA.pdf

Lloyd SM, Ries R (2007) Characterizing, propagating, and analyzing uncertainty in life-cycle assessment: a survey of quantitative approaches. J Indust Ecol 11(1):161–179

Weidema BP (1998) Multi-user test of the data quality matrix for product life cycle inventory data. Int J Life Cycle Assess 3 (5):259–265

Weidema BP, Wesnæs MS (1996) Data quality management for life cycle inventories—an example of using data quality indicators. J Clean Prod 4(3–4):167–174

Weidema BP, Bauer C, Hischier R, Mutel C, Nemecek T, Vadenbo CO, Wernet G (2011) Overview and methodology: data quality guideline for the ecoinvent database version 3 (final draft_revision 1) ecoinvent report No. 1(v3), http://www.ecoinvent.org/fileadmin/documents/en/ecoinvent_v3_elements/01_DataQualityGuideline_FinalDraft_rev1.pdf